D-025

階層クラスタリングにおける新しい手法の提案

Proposal of a new algorithm for hierarchical clustering

林　達彦†　　小柳　滋†
Tatsuhiko Hayashi　Shigeru Oyanagi

1．はじめに

代表的なデータ解析手法であるクラスタリングには，最短距離法などの階層的手法と，K-means などの分割最適化手法に分けられることができるが，どの手法に注目してみても何かしらの問題点を持っている．本研究では凝集型の階層的クラスタリング手法に注目する．

凝集的な階層的クラスタリングには 6 つの代表的な手法（最短距離法，最長距離法，群平均法，重心法，メディアン法，ウォード法）が挙げられるが，この中に常に最適な解を得る手法があるのではなく，扱うデータセットによって探索的にクラスタリングを繰り返し最適な解を求めていくことになる．

本研究では，最短距離，最長距離，重心，メディアンの 4 つの手法のアルゴリズムの矛盾に注目しつつ，ウォード法や群平均法よりも計算時間が小さい，新たな 7 つ目の手法を提案する．

2．既存手法

2.1　既存手法の欠点

凝集型の階層的クラスタリングにおける既存の 6 つの手法とその欠点について述べる．

これらの手法では，全データをそれぞれ 1 つのクラスタであると考える．そして全てのクラスタの組に対して距離（非類似度）を求め，クラスタ間距離が最小なクラスタの組を統合し新たな 1 クラスタとする．そして全てのデータが 1 クラスタに統合されるまでクラスタの統合を繰り返し，全過程がデンドログラムとして表される．ここでクラスタ間の距離をどのように定義するか，すなわちどのような 2 クラスタを近いクラスタとするか，ここに各手法の違いがある．

最短距離法と最長距離法は、2 つのクラスタ間の最も小さい距離となる 2 データ，もしくは，最も大きい距離となる 2 データの組み合わせをクラスタ間の距離として定義する．最短距離法はより多くのデータが統合されたクラスタが他のクラスタを統合しやすい性質があり，逆に最長距離法は統合しにくい性質があることが容易に分かる．単にデータ間を連結させたいときには同じアルゴリズムで単連結法，完全連結法と呼ばれる手法が役に立つが，クラスタリングとは単にデータを結合することが目的でなく，類似したデータを同じクラスタに，類似しないデータを別のクラスタにまとめる作業であるので，最短距離法と最長距離法はアルゴリズムに矛盾を持った手法であることは否めない．

最短距離法や最長距離法のようにクラスタ間距離を極端なデータ間距離の値とするのではなく，全データ間の距離をクラスタ間距離に反映させるのが群平均法である．

† 立命館大学大学院

対象の 2 クラスタ内のデータ全ての組み合わせの距離の平均値をクラスタ間距離にする群平均法なら，クラスタ間距離の定義，比較に矛盾を起こしにくい．データ数やデータあたりの要素数が増えれば計算時間が膨大になってしまうことが欠点になる．

一方で重心法とメディアン法では，各クラスタの位置をクラスタ内のデータの重心やメディアンとして定め，そこから距離を計算する．最短距離法，最長距離法に比べると計算時間は大きくなるものの，群平均法，ウォード法と比べると小さい．しかし群平均法と比べて距離計算が単純であるため，良い結果は生みにくくなる．

最後にウォード法だが，これは他の手法とは異なり，クラスタ間の距離を考えるのではなく，クラスタの分散をできるだけ小さくするようにデータを統合していくアルゴリズムである．最も適確なデンドログラムを生成する手法であると言われている．多くの場合でこの手法を使うのが望ましいが，データ数や次元数が大きくなるとやはり計算時間が膨大になることも知られている．

2.2　既存手法まとめ

クラスタの分散を考慮するウォード法と，クラスタ内データの全組み合わせの距離を計算する群平均法は、他の手法に比べて矛盾のないアルゴリズムである一方で計算時間というネックを持っている。

他の 4 つの手法の計算時間は短いが，アルゴリズムに矛盾を抱えている．その矛盾とはどんな理由から起きてしまうのか，群平均法と比べて手抜きしているところを考えてみると分かる．図 1 に示すように，複数のデータを統合したクラスタは当然その範囲が広がっているにも関わらず，ある一点からある一点までの距離をクラスタ間距離としていることに矛盾を起こす可能性を秘めている．

次節ではウォード法や群平均法より計算時間が小さく，従来手法には無い概念を持った手法を提案する．

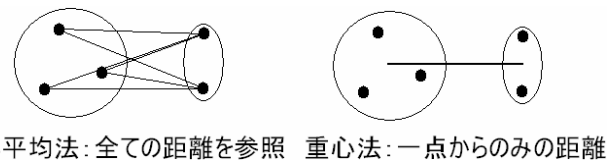


図 1：群平均と重心法の比較

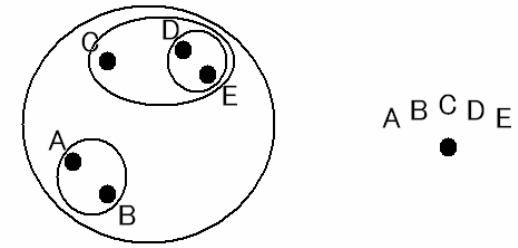
3．提案する手法

3.1　データの仮想的移動

この手法ではクラスタの範囲を広げて複数のデータを統合させるのではなく，クラスタそのものをある一点上に存在させる．そのため，同じクラスタに属するデータは全て一点上に重なって存在する．従来手法ではデータの位置はあくまで固定されていて，それらを囲うクラスタが範囲を

広げていったが，提案する手法では統合されるべき 2 データが互いの中点となる位置に仮想的に移動する．単純に 2 データが中点となる座標上に移動するのではなく，まるで 2 データ間の空間が切り取られるように周囲のデータも同じベクトル分の影響を受け移動する．

従来手法では全てのデータが一つの大きなクラスタに囲まれるとクラスタリング完了だったが，本手法では全てのデータがある一点上に重なるとクラスタリング完了となる（図 2）．



従来手法の終了イメージ　提案手法の終了イメージ

図 2：クラスタリング完了時のイメージの違い

3.2　第一ステップの反復

従来のどの手法においても必ず共通のクラスタ統合が行われるステップがあり，それが最初のステップである．全てのデータが別々のクラスタと見なされている初期状態では，最も小さい距離に配置されている 2 データが 1 つのクラスタに統合される．つまり「全てのデータの中で最も近い組み合わせの 2 データは，最初にクラスタリングされるべき 2 データである」という考えに矛盾が無いことが分かる．すなわち，もしこの矛盾の無い 1 ステップ目を反復して行うことができれば理想的なクラスタリングが行えるのではないかと本研究では予想する．

本研究で提案する手法は統合された 2 データが 1 点に重なるので，2 ステップ以降もまるで従来手法の 1 ステップ目のようにデータが存在する（またデータ数は一つ減った状態になる）．すなわち 2 ステップ目以降も同じように最も近い 2 データを探して統合すればよい．

このようにデータを仮想的に移動するという従来手法には無い概念を取り入れることによって，従来手法全てに共通した最初のステップを反復することができる．

4. 提案手法の検証

4.1　計算時間について

既存の 6 つの手法の中で計算時間の大きさがネックとなっているのがウォード法と群平均法のクラスタ間距離の計算である．ここでは提案する手法が 2 つの手法よりクラスタ間距離の計算時間が小さくなっているか検証する．

n 個のデータを含むクラスタと m 個のデータを含むクラスタ間の距離を計算すると仮定する．群平均法では(n*m)回の距離計算を行い，その平均値をクラスタ間距離としている．ウォード法では 2 クラスタの持つ全データの重心を計算し，(n+m)回の距離計算を行う．以上のことからクラスタサイズが大きくなればなるほど距離計算の時間が膨大になることは自明である．

提案する手法ではクラスタそのものがある一点上に存在しているという考え方であるため，1 クラスタに含まれるデータの数に関係なく，クラスタ間の距離は 1 回の距離計算で完了する．これは他の 4 つの手法と共通している．データの仮想移動において全データの座標の更新が必要になるが，クラスタの統合が進むにつれてこの回数は減少していくので最低で 1 回，最高でも全データ数に等しい回数の更新になる．

4.2　生成するクラスタリングについて

クラスタ内の極端なデータを参照する最短距離法と最長距離法は，本手法とはたびたび違うクラスタ生成をすることは容易に予想される．そこで計算時間の小さい残りの手法の重心法，メディアン法と違うクラスタ統合を行う可能性があることをここでは示す．

図 3 は，2 次元平面上に 4 つのデータ a,b,c,d が配置されていたと仮定し，互いの距離関係を表したものである．全ての手法の最初のステップにおいて，クラスタ統合されるデータは a と b である．ここで，データ a,b が統合されたクラスタとデータ d では，どちらがデータ c に近いだろうか．重心法とメディアン法においては a,b と c の距離は,a,b の中点と c の距離を計算するため,距離は 12 となり,cd 間の距離 12.5 より近くなる．提案手法においては a,b と c 間の距離はやはり 12 と計算されるのだが，cd 間の距離が 12 未満になると考えられる（d が仮想的に移動して c に近づくため）．

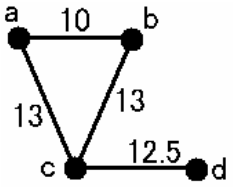


図 3：データ配置の例

5. おわりに

先に述べたように，クラスタリングではどの手法が良い結果を生むかという優劣を比べるものではない．一つのデータセットに対して様々な手法で探索的にクラスタリングを行うのが理想的である．しかしながら，群平均法やウォード法の計算時間が気になるときに代用の手法が既存の 4 つでは足りないと考え，本研究では新しい手法を提案した．この手法は既存手法には無い概念を取り入れることで，既存手法全てに共通する最初のステップを反復するという新たな試みを持っている．

この手法を用いたときに他の手法では生成されないクラスタリングがみられる可能性に期待し，今後はこの手法が持つ特徴を研究していく．また同時に，クラスタリング分野特有の問題である順序依存性や次元の呪いなどの影響についても調べ，可能な限りアルゴリズムの向上に努めたい．

参考文献

神嶌 敏弘, データマイニング分野のクラスタリング手法 (1) ,(2), 人工知能学会誌, vol.18, no.1, pp.59-65, no.2, pp.170-176 (2003)